

protoDUNE prompt processing system (p3s): status and plans

Maxim Potekhin, Brett Viren (BNL)

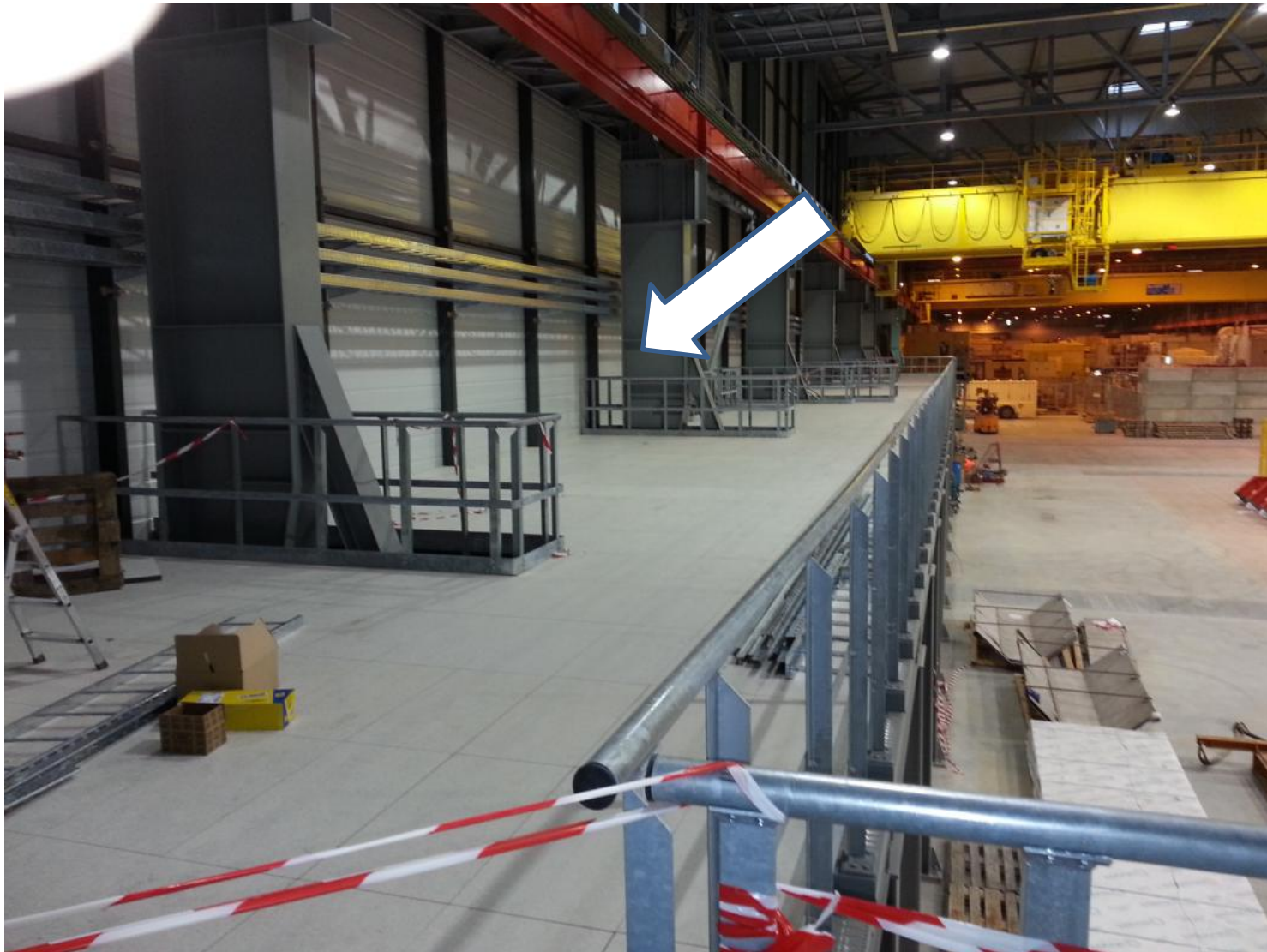
DUNE Weekly @ BNL

12/07/2016

Recent news and general status of NP04 computing

- protoDUNE DAQ review has taken place at CERN, materials available online
- Shortage of manpower in monitoring and slow controls has been noted
- This year's FNAL computing budget has substantially shrunk. Implications for storage and other computing resources for protoDUNE. Projections out to 2018 are TBD.
- This is a challenge and an opportunity for other labs to assume a bigger role
 - will restart consulting with BNL RACF regarding opportunistic resources and available storage
 - may motivate direct replication from CERN to BNL in addition to FNAL
- No significant new development in protoDUNE data handling (FNAL responsibility) and nothing substantially new in the DAQ area
- Computing retreat (by invitation) at FNAL on Dec 6th-7th. TBD:
 - org issues
 - strategy out to 2020
- A large part (or all) of the Neutrino Platform cluster ("neut") will be located in EHN1 in a water cooled enclosure. This is apparently intended for both NP02 and NP04, with partitioning. Physical access from the control room seems not too inconvenient (next slide).

Space for control rooms in EHN1



The other side of EHN1 (data rooms location)



Brief History of p3s

- From the start, Data Quality Management (DQM) and the prompt processing system are a part of the protoDUNE Computing Model .
- In September 2016 the interface between the DAQ and the “offline” was defined as the boundary of the online buffer (DELL storage appliance and servers). No longer responsibility of BNL.
- In the Spring of 2016 the CERN Neutrino Platform organization acquired a few dozen of older compute server units (DELL PowerEdge) which are still servicable, with plans to increase their number to ~300. These nodes feature 8 cores and 16GB of RAM and are currently in light-duty use for MC, reco and testing purposes. This cluster has been named “neut” (short for Neutrino Platform). We realized that it can be extremely useful for DQM and can support a large portion of prompt processing. The part of the cluster used for DQM will be named "neutdqm".
- The outline of the design of the **protoDUNE prompt processing system** (now known as **p3s**) has been documented in DocDB **1861** which is actively maintained. Please refer to that document for details.

Goals

- According to the Requirements (DocDB1811) the data needs to undergo prompt processing on the scale of “tens of minutes” (or better) from the time it was taken, with the benchmark number of 10 min often quoted for reference purposes. This is motivated by the need to have *actionable* DQM information in time for operators to take action and prevent loss of useable data and/or valuable beam time.
- The goal is to provide a more in-depth data QA and assessment of the detector's health and operating conditions than is afforded by the generic monitoring running as artDAQ processes on the DAQ machines.
- Most calculations done in p3s will result in some sort of a “visual product”, for example a histogram, plot, event display or a summary table in order to provide information to the experiment operators in a timely manner.
- Since the volume of visual and numeric information pertinent to DQM may be large and hard to quickly comprehend, there will be automated alerts to the operators when certain parameters are outside of their nominal range.

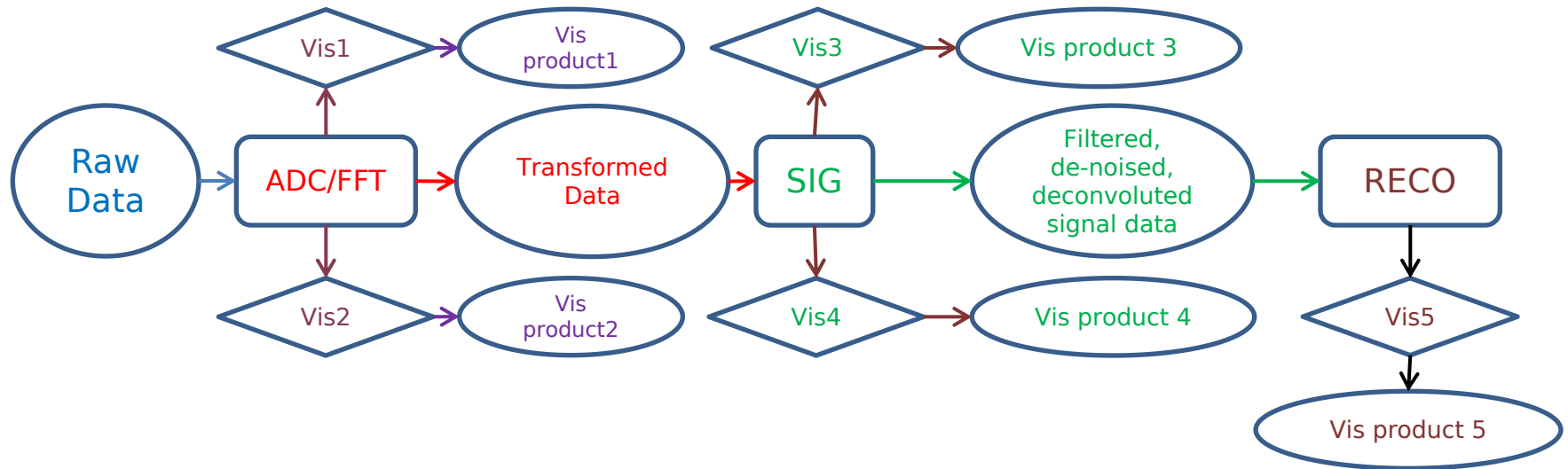
Categories of the TPC data processing in p3s

- **DAQ** (no data decompression): A summary of DAQ-level data such as summaries of data rates or summaries of any metadata, status codes the provided by the DAQ etc. Parts of this functionality will exist within the artDAQ monitor processes, perhaps with additional detail provided by the prompt processing system
 - data propagates within DAQ mostly in compressed form
- **ADC** (requires data decompression): A summary of ADC-level data e.g. mean/RMS values at channel level and as a statistics over various groupings and level of detail (ASIC, FEMB, RCE, APA etc).
- **FFT**: A summary of the ADC-level data in frequency space. It requires running a discrete Fourier transform (FFT) on channel waveforms. This largely provides measures of noise and its evolution.

Categories of the TPC data processing in p3s (cont'd)

- **SIG:** A summary of the data after signal processing. The processing is in both time domain and in frequency domain (so uses the output of FFT). It includes items such as
 - “stuck code” mitigation
 - coherent noise removal
 - noise subtraction and filtering
 - deconvolution of the response function
 - calculation of signal correlations for diagnostic purposes
- **RECO:** Results from running some type of reconstruction (perhaps greatly simplified). It may, for example, provide a coarse count of straight muon track candidates. We are working with the protoDUNE reco leaders (Robert and Dorota) to better define the scope and other detail.
- Examples:
 - channel vs time plots before and after **SIG**
 - track fitter working on raw data

Prompt Processing as a DAG (concept)



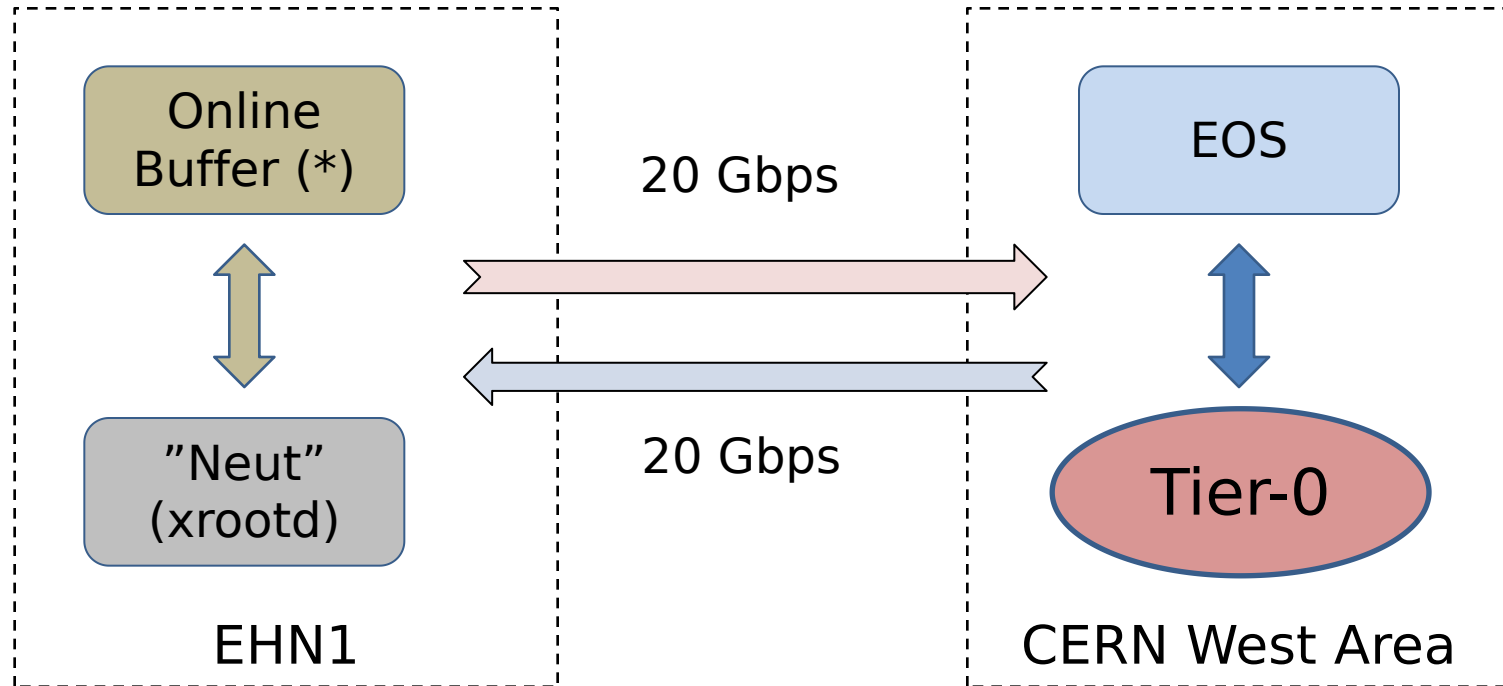
Automation

- Automation at more than one level is obviously crucial for p3s to be of value.
- Processing is triggered by arrival of fresh raw data in the first stage, and then by arrival of processed data from the previous stage as the dataflow progresses.
- Automatic adjustment of priorities and preemption to free up resources as needed.
- Those data products which are designated for archival (for later audit/diagnostics etc) need to be committed to long-term storage automatically.
- Alarms.

Prioritization and prescaling of data

- Processing the full data stream won't be possible under any realistic allocation of resources, so the data must be pre-scaled/sampled in each stage of processing, according to a policy based on the desired frequency of production of certain visual and data products for presentation to protoDUNE operators
 - cf. how often do we need to produce a crude event display? RMS strip charts?
- Reading only a part of the data from a given file (e.g. a single readout frame) is crucial for minimizing the additional I/O load on storage/network; xrootd can give us that and this must be exploited.
- Fraction of the data which “makes it” through the chain is not fixed and may be adjusted dynamically (including automatically), jobs can be dropped to free up resources to satisfy peak demand in a certain stage etc.
- All of these factors create a substantial difference between p3s and a typical WMS such as used on the Grid. It's less deterministic and loss of jobs and data can be accepted while still actively managed.

Options for data access in p3s



*NB. There is a preliminary agreement with the DAQ team that the Online Buffer will run xrootd server software which is low-weight and will be mostly idling (see next slide).

EOS vs the Online Buffer: the switch-over

- Relative placement of the computing resource (e.g. neut) vs data and from what source the data is fed to p3s is still a matter of debate
- Nominal current design: p3s will utilize “neut” in EHN1 and there will be functionality provided for bimodal operation
 - reading the data from EOS under normal conditions
 - reading from the Online Buffer using preferably xrootd when EOS and/or the data link to central campus becomes unavailable
- This still needs more understanding as transmission of data to EOS introduces extra latency and the aim is to keep every step at less than a minute scale
 - recently highlighted by T.Junk
- Again, the xrootd interfaces of both EOS and the Online Buffer provide commonality of data access which simplify such tactics.

Resources

- Utilization of the outermost layer of DAQ to do DQM as a part of its own monitoring is TBD (i.e. the scope and volume)
- The “neutdqm cluster”:
 - adequate rack space + cooling is foreseen in EHN1 on the side of the hall opposite the control rooms
 - a few machines can be used for creating services for the workflow orchestration
- Hardware will also be needed at FNAL to mirror the database and web service for optimal performance and scalability.
- CERN Tier-0 at the scale of ~ 1000 cores (TBD)
 - remains to be decided what type and fraction of prompt processing can run/needs to run in Tier-0
- It is very efficient to have a single system in place which manages both pools of resources and perhaps some resources outside of CERN, so this design parameter is built into p3s

p3s: the design parameters

- We need a simple, portable and lightweight real time workflow management system to operate in the environment described in previous slides.
- It's different from popular WMSs according to a few criteria noted above
- Design parameters:
 - Do not develop/deploy a data handling system. Instead, leverage xrootd
 - Do use a database to keep track of the state of jobs in various categories.
 - Do not rely on a particular flavor of batch system in the design since there are at least two different ones at CERN (cf. LSF and HTCondor).
 - Do take measures to counter the latency and lack of deterministic job submission in order to make the processing happen “just in time”.
 - Do build a presentation layer for the visual and other products to be made available to the protoDUNE operators.
 - Create a “clickable” and interactive UI to be able to quickly navigate around the system and details of its state for diagnostics and management purposes.

p3s: the system

- The design parameters listed above can be met by a pilot-based system with a database backend.
- This effectively results in a design based on the web service approach (pilots communicating with the central service and jobs being dispatched to pilots)
- Choices for the web service framework are many. Due to existing in-house expertise and ease of use we selected Django as the platform.

p3s: components

- Pilot Factory. Software reuse — may utilize the AutoPyFactory supported as a project by DOE, previously used in ATLAS at scale. It's system-agnostic and can be adapted for protoDUNE.
- Job Generator. Triggered by the arrival of input data subject to configurable criteria. Can we reuse a module or two from F-FTS? Jobs are created as entries in the queue (implemented as a DB table) waiting to be matched with an available WN slot.
- The Server. It's the centerpiece of the system and dispatches jobs to active and validated pilots.
- The Monitor. Provides the view of the system including the Jobs, Pilots and state of the data.
- Presentation layer. Serves the results to the user.








p3s: development

- Development of p3s started a few weeks ago.
- The first rough prototype is ready, with a few components listed above just at the mockup stage and are being fleshed out gradually.
- Classes of objects in p3s:
 - jobs
 - pilots
 - workflows
 - data products
- Priority policy mechanism is implemented, configurable in the DB.
- Simple and minimalistic code.
- Leveraging third-party packages to create tables etc with no coding to speak of.
- Repo on GitHub. Enthusiasts/volunteers/tinkerers are welcome to join.
- Ready to start migration to PostgreSQL and “in situ” testing at CERN.
- Ongoing communication with both DAQ and Reco team to move towards well understood interfaces and deliverables.

p3s: a few screenshots

pilots

Current time: 11/29/16 18:49:21

ID 	state 	site 	host 	ts_cre 	ts_reg 	ts_lhb 
1	active	default	ferocity	11/29/2016 4:10 p.m.	11/29/2016 4:10 p.m.	11/29/2016 4:10 p.m.
2	active	default	serenity	11/29/2016 5:01 p.m.	11/29/2016 5:01 p.m.	11/29/2016 5:01 p.m.
3	active	default	serenity	11/29/2016 5:27 p.m.	11/29/2016 5:27 p.m.	11/29/2016 5:27 p.m.
4	failed brokerage	default	serenity	11/29/2016 6:20 p.m.	11/29/2016 6:20 p.m.	11/29/2016 6:20 p.m.
5	failed brokerage	default	serenity	11/29/2016 6:37 p.m.	11/29/2016 6:37 p.m.	11/29/2016 6:37 p.m.
6	failed brokerage	default	serenity	11/29/2016 6:38 p.m.	11/29/2016 6:38 p.m.	11/29/2016 6:38 p.m.
7	failed brokerage	default	serenity	11/29/2016 6:45 p.m.	11/29/2016 6:45 p.m.	11/29/2016 6:45 p.m.
8	failed brokerage	default	serenity	11/29/2016 6:47 p.m.	11/29/2016 6:47 p.m.	11/29/2016 6:47 p.m.
9	failed brokerage	default	serenity	11/29/2016 6:47 p.m.	11/29/2016 6:47 p.m.	11/29/2016 6:47 p.m.
10	failed brokerage	default	ferocity	11/29/2016 6:48 p.m.	11/29/2016 6:48 p.m.	11/29/2016 6:48 p.m.
11	failed brokerage	default	ferocity	11/29/2016 6:48 p.m.	11/29/2016 6:48 p.m.	11/29/2016 6:48 p.m.
12	failed brokerage	default	ferocity	11/29/2016 6:48 p.m.	11/29/2016 6:48 p.m.	11/29/2016 6:48 p.m.
12 pilots						

p3s: a few screenshots

[Home](#)

P3S Jobs

Current time: 11/27/16 16:58:30

ID	uuid	p uuid	stage	priority	state	ts_def	ts_dis	ts_sta	ts_sto
83	39d0f176-b1e4-11e6-a783-0022693c6a17	—	stage1	1	defined	11/23/2016 8:20 p.m.	—	—	—
84	39d0f177-b1e4-11e6-a783-0022693c6a17	—	stage2	2	defined	11/23/2016 8:20 p.m.	—	—	—
85	39d0f178-b1e4-11e6-a783-0022693c6a17	—	stage3	3	defined	11/23/2016 8:20 p.m.	—	—	—
86	39d0f179-b1e4-11e6-a783-0022693c6a17	—	stage4	4	defined	11/23/2016 8:20 p.m.	—	—	—
87	39d0f17a-b1e4-11e6-a783-0022693c6a17	—	stage5	5	defined	11/23/2016 8:20 p.m.	—	—	—
88	39d0f17b-b1e4-11e6-a783-0022693c6a17	bb08d790-b1e4-11e6-a783-0022693c6a17	stage6	6	finished	11/23/2016 8:20 p.m.	11/23/2016 8:24 p.m.	—	—
89	39d0f17c-b1e4-11e6-a783-0022693c6a17	6b549946-b1e4-11e6-a783-0022693c6a17	stage7	7	finished	11/23/2016 8:20 p.m.	11/23/2016 8:22 p.m.	—	—
90	39d0f17d-b1e4-11e6-a783-0022693c6a17	6b3ff022-b1e4-11e6-a783-0022693c6a17	stage8	8	finished	11/23/2016 8:20 p.m.	11/23/2016 8:22 p.m.	—	—
91	46591ebe-b1e4-11e6-a783-0022693c6a17	—	stage1	1	defined	11/23/2016 8:21 p.m.	—	—	—
92	46591ebf-b1e4-11e6-a783-0022693c6a17	—	stage2	2	defined	11/23/2016 8:21 p.m.	—	—	—
93	46591ec0-b1e4-11e6-a783-0022693c6a17	—	stage3	3	defined	11/23/2016 8:21 p.m.	—	—	—
94	46591ec1-b1e4-11e6-a783-0022693c6a17	—	stage4	4	defined	11/23/2016 8:21 p.m.	—	—	—
95	46591ec2-b1e4-11e6-a783-0022693c6a17	—	stage5	5	defined	11/23/2016 8:21 p.m.	—	—	—
96	46591ec3-b1e4-11e6-a783-0022693c6a17	b68488d4-b358-11e6-a72a-0022693c6a17	stage6	6	finished	11/23/2016 8:21 p.m.	11/25/2016 4:47 p.m.	—	—
97	46591ec4-b1e4-11e6-a783-0022693c6a17	bb0f65f6-b1e4-11e6-a783-0022693c6a17	stage7	7	finished	11/23/2016 8:21 p.m.	11/23/2016 8:24 p.m.	—	—
98	46591ec5-b1e4-11e6-a783-0022693c6a17	6b50a598-b1e4-11e6-a783-0022693c6a17	stage8	8	finished	11/23/2016 8:21 p.m.	11/23/2016 8:22 p.m.	—	—
99	521a2716-b1e4-11e6-a783-0022693c6a17	—	stage1	1	defined	11/23/2016 8:21 p.m.	—	—	—
100	521a2717-b1e4-11e6-a783-0022693c6a17	—	stage2	2	defined	11/23/2016 8:21 p.m.	—	—	—
101	521a2718-b1e4-11e6-a783-0022693c6a17	—	stage3	3	defined	11/23/2016 8:21 p.m.	—	—	—
102	521a2719-b1e4-11e6-a783-0022693c6a17	—	stage4	4	defined	11/23/2016 8:21 p.m.	—	—	—
103	521a271a-b1e4-11e6-a783-0022693c6a17	—	stage5	5	defined	11/23/2016 8:21 p.m.	—	—	—
104	521a271b-b1e4-11e6-a783-0022693c6a17	—	stage6	6	defined	11/23/2016 8:21 p.m.	—	—	—
105	521a271c-b1e4-11e6-a783-0022693c6a17	bb11de8a-b1e4-11e6-a783-0022693c6a17	stage7	7	finished	11/23/2016 8:21 p.m.	11/23/2016 8:24 p.m.	—	—
106	521a271d-b1e4-11e6-a783-0022693c6a17	6b4feb26-b1e4-11e6-a783-0022693c6a17	stage8	8	finished	11/23/2016 8:21 p.m.	11/23/2016 8:22 p.m.	—	—

24 jobs

p3s: a few screenshots

[Home](#)

P3S Jobs

Current time: 11/28/16 21:20:43

Attribute ▲	Value ▲
id	89
p_uuid	6b549946-b1e4-11e6-a783-0022693c6a17
priority	7
stage	stage7
state	finished
ts_def	Nov. 23, 2016, 8:20 p.m.
ts_dis	Nov. 23, 2016, 8:22 p.m.
ts_sta	—
ts_sto	—
uuid	39d0f17c-b1e4-11e6-a783-0022693c6a17
10 items	

p3s: people and plans

- [Nektarios] Lots of infrastructure work still needs to be done on the CERN cluster
 - xrootd management
 - Apache
 - PostgreSQL
 - Pilot Factory
- [Maxim] Development of p3s is ongoing and appears to be on track
- [Brett] Formats and interfaces are yet to be established with DAQ and reco groups